THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 1, January 1983
Printed in U.S.A.

Stochastic Analysis of Mechanizing Transaction Data Bases

By J. A. MORRISON and W. W. YALE

(Manuscript received February 16, 1982)

In this paper we consider the transfer of records from a manual file system (MFS) to a mechanized data base management system when the conversion takes place through two processes. In the going-forward process a transfer is made by the regular work force when an order is received to change or delete a record in the MFS. In addition, the transfer of records is carried out by a crash force, working at a fixed rate. We investigate the expected number of records remaining in the MFS at future times, and the expected number of records removed from the MFS during the corresponding periods by the going-forward work force and by the crash force. We also derive the expected time taken to go from a prescribed number of records in the MFS to a smaller number. We give simple approximations for all of these quantities. The results have been used elsewhere in the construction of an economic model used to estimate cost/benefits and labor force levels during the mechanization of transaction data bases. The numerical results presented graphically are typical of two Bell System applications.

I. INTRODUCTION

In this paper, we analyze the generic problem of implementing a system that requires manual effort for converting paper records in a manual file system (MFS) to mechanized computer records in a data base management system (DBMS). We can achieve conversion to a mechanized environment in a number of ways: by going-forward on an activity basis, by crash-conversion utilizing an augmented work force, or by some mix of the two dictated by available resources such as money, people, space, and time objectives.

A paper record is mechanized using the going-forward approach if the record is transferred from the MFS to the DBMS by regular record

maintenance employees when a change or deletion activity is performed upon that record. If a randomly chosen record is transferred by a separate temporary work force, the record is said to be transferred using the crash approach. The paper record could contain all of the information that is intended to be mechanized or an enhancement to an already existing mechanized record.

The arrival times of activities (change or deletion orders) for a particular record are assumed to form a Poisson process. We further assume that the distributions of the activity arrival times for every record are independent and identical. Hence, the sequence of times at which these change and deletion activities occur forms a nonstationary Poisson process. In addition, the transfer of records is carried out by a special task force with constant total mean rate, the amount of work required for a record to be transferred being exponentially distributed. Hence, there are two different stochastic processes that are concurrently operating against the MFS. Each record in the MFS is transferred if an activity occurs upon it, while at the same time each record in the MFS has a chance of randomly being selected by a crash person when he/she completes a conversion.

When an activity occurs against a record in the MFS, it will be assumed that a going-forward person will immediately attend to transferring and processing the record. Therefore, if substantially more than the average number of records concurrently experience activity, it will be assumed that management will borrow going-forward people from other areas or have the normal going-forward force work overtime. However, the time it takes for a crash-conversion person to perform the conversion is pertinent since the force is assumed to be fixed. In this case, the average number of records that the entire crash conversion force can convert in a specified period determines the mean of the stochastic process describing the crash effort.

In this paper we highlight the derivations of formulas that are used in the construction of a model[1] used to estimate cost/benefits and labor force levels during the mechanization of transaction data bases. In the next section, we investigate the expected number of records remaining in the MFS, removed by the going-forward people, and removed by the crash force people. In particular, we point out a simple approximation to the expected number of records remaining in the MFS, which is valid over a significant part of the range of interest. We derive the expected passage time from l to m records in the MFS ($l > m$) in Section III, and obtain a simple asymptotic approximation for the expected time at which the MFS becomes empty. Finally, in Section IV, we present some conclusions, and discuss the relevance of the results to Bell System applications.

II. EXPECTED NUMBER OF RECORDS

We will now consider a stochastic model for the transfer of records from a MFS to a DBMS. We assume that there are n crash force people, and that the conversion times of the records are independent and exponentially distributed with mean rate ρ. At time $\tau = 0-$ there are S records in the MFS. At time $\tau = 0$ there are n of these records removed from the MFS by the crash force. When the conversion of a record is completed, that record is instantaneously transferred to the DBMS, and another record is removed from the MFS by the crash person involved. A record is also removed from the MFS, by a going-forward person, when a change or deletion order is received; that record is transferred to the DBMS at some later time. We assume that the sequence of times at which change and deletion orders are received forms a nonstationary Poisson process with rate $i\mu$ where i is the number of records in the MFS. (The mean deletion and change order rates are summed to get μ, since it is a well-known fact that if two independent Poisson processes are affecting the records, then the composite of these two processes can be described by a Poisson process where the parameters are summed.) If a change or deletion order is received for a record that is in the process of being converted by a crash person, the change or deletion order is not executed until after the record has been transferred to the DBMS.

The number of records in the MFS at time $\tau = 0+$ is

$$M = S - n. \tag{1}$$

For convenience, we let

$$n\rho = \nu\mu. \tag{2}$$

This is the mean rate at which the crash force converts records. We also let $P_i(\tau)$ denote the probability that there are i records in the MFS at time τ. But, the number of records in the MFS is a pure death process,[2] with death rate

$$\mu_i = n\rho(1 - \delta_{i0}) + i\mu = \mu[\nu(1 - \delta_{i0}) + i], \qquad i = 0, \cdots, M, \tag{3}$$

where δ_{ij} denotes the Kronecker delta; $\delta_{ij} = 1$ if $i = j$, and $\delta_{ij} = 0$ if $i \neq j$. Hence,

$$\frac{dP_i}{d\tau} = -\mu[\nu(1 - \delta_{i0}) + i]P_i + \mu(1 - \delta_{iM})(\nu + i + 1)P_{i+1}, \tag{4}$$

for $\tau > 0$ and $i = 0, \cdots, M$. The initial condition at time $\tau = 0+$ is

$$P_i(0+) = \delta_{iM}, \qquad i = 0, \cdots, M. \tag{5}$$

The expected number of records in the MFS at time τ is

$$N(\tau) = \sum_{i=1}^{M} iP_i(\tau). \tag{6}$$

It follows from (4) through (6) that

$$\frac{dN}{d\tau} + \mu N = -\nu\mu[1 - P_0(\tau)]; \qquad N(0+) = M. \tag{7}$$

It is straightforward to calculate the Laplace transforms of P_i, $i = 0, \cdots, M$, from (4) and (5), and thence that of N from (7). By inverting the latter transform, Morrison[3] obtained an explicit expression for $N(\tau)$ in terms of an incomplete beta function.[4] Asymptotic approximations to $N(\tau)$, involving the complementary error function,[4] were derived when $M \gg 1$ and $\nu \gg 1$.

Since $0 \le P_0(\tau) \le 1$, it follows from (7) that

$$L(\tau) \equiv (M + \nu)e^{-\mu\tau} - \nu \le N(\tau) \le Me^{-\mu\tau}. \tag{8}$$

The lower bound is, of course, superfluous if $L(\tau) \le 0$. When $M \gg 1$ and $\nu \ll M$, these bounds yield an accurate approximation to $N(\tau)$ as long as $Me^{-\mu\tau} \gg \nu$. An upper bound for $N(\tau)$, which is valid only for $L(\tau) \ge 0$, and exact for $L(\tau) = 0$, was derived from the exact result.[3] It was shown that

$$0 \le N(\tau) - L(\tau) \le \frac{\Gamma(M + \nu)}{\Gamma(M)\Gamma(\nu)} e^{-\nu\mu\tau}(1 - e^{-\mu\tau})^M \quad \text{for} \quad L(\tau) \ge 0. \tag{9}$$

From (9) we deduce that

$$N(\tau) \approx L(\tau) \quad \text{for} \quad M \gg 1, \qquad L(\tau) \gg \min(\sqrt{M}, \sqrt{\nu}) \frac{\nu^\nu e^{-\nu}}{\sqrt{\nu}\Gamma(\nu)}, \tag{10}$$

where, from the properties of the gamma function,[4]

$$\frac{\nu^\nu e^{-\nu}}{\Gamma(\nu)} < \nu \quad \text{for} \quad \nu > 0; \qquad \frac{\nu^\nu e^{-\nu}}{\sqrt{\nu}\Gamma(\nu)} \sim \frac{1}{\sqrt{2\pi}} \quad \text{for} \quad \nu \gg 1. \tag{11}$$

Hence, (10) gives a larger range of validity for the approximation $N(\tau) \approx L(\tau)$ than does (8), and significantly so if $\nu \gg 1$, as it is in cases of interest. We remark that intuitively we expect $P_0(\tau)$, the probability that there are no records in the MFS at time τ, to be extremely small for a significant time when $M \gg 1$, and in that time interval we deduce from (7) that $N(\tau) \approx L(\tau)$. The intuitive derivation, however, does not give the range of validity of the approximation.

Typical numerical results are illustrated in Figs. 1 and 2 for the case $S = 10^5$, $n = 10$, $\mu = 0.001608$ per day, and $\nu = 18657$. Only $N(\tau)$ and the lower bound $L(\tau)$ are shown in Fig. 1. The upper bound given by (9) is also shown in Fig. 2. This figure depicts the tail region, in which the expected number of records in the MFS is considerably fewer than

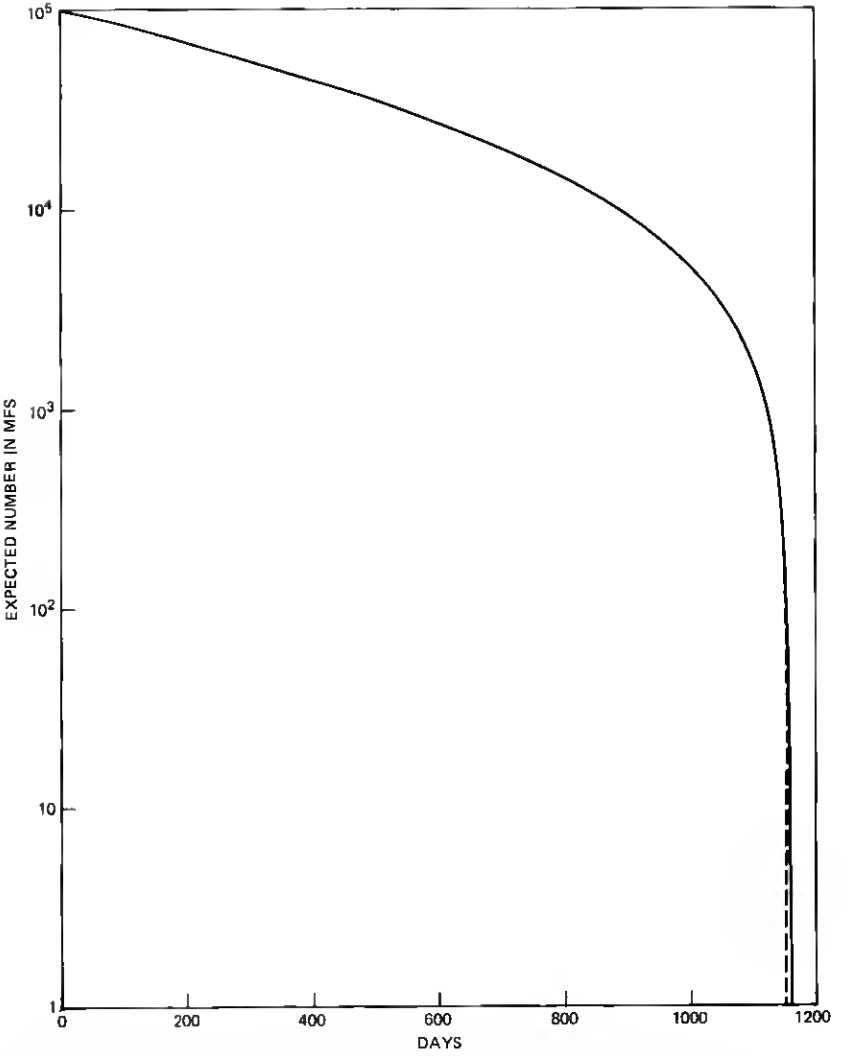

Fig. 1—Expected number of records in the MFS, and lower bound, as a function of time, for $S = 10^6$, $n = 10$, $\mu = 0.001608$ per day, and $\nu = 18657$.

the initial number. It is seen that $N(\tau)$ is still quite close to $L(\tau)$ for values of $L(\tau)$ just a few times greater than $\min(\sqrt{M}, \sqrt{\nu})/\sqrt{2\pi} \approx 54.5$.

Another quantity of interest is the expected number of records $F(\tau)$ removed from the MFS by the going-forward people in the time interval $(0, \tau]$. Let $P_{ik}(\tau)$ denote the probability that there are i records in the MFS at time τ, and that k records have been removed from the MFS by the going-forward people in $(0, \tau]$. Then

$$F(\tau) = \sum_{i=0}^{M-1} \sum_{k=1}^{M-i} kP_{ik}(\tau). \tag{12}$$

But,

$$\frac{dP_{ik}}{d\tau} = -\mu[\nu(1 - \delta_{i0}) + i]P_{ik} + \mu\nu(1 - \delta_{k,M-i})P_{i+1,k} + \mu(1 - \delta_{k0})(i + 1)P_{i+1,k-1}, \tag{13}$$

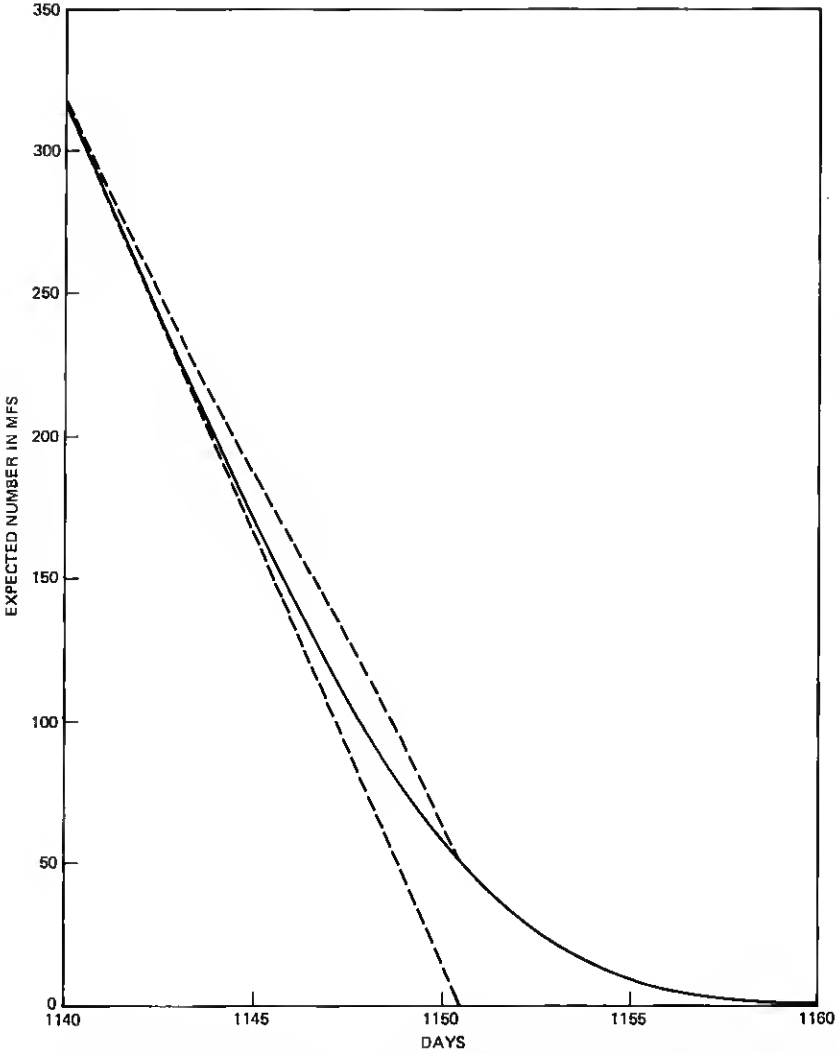

Fig. 2—Expected number of records in the MFS, and upper and lower bounds, as a function of time, for $S = 10^5$, $n = 10$, $\mu = 0.001608$ per day, and $\nu = 18657$.

for $\tau > 0$, $k = 0, \cdots, M - i$, and $i = 0, \cdots, M$. The initial condition at time $\tau = 0+$ is

$$P_{ik}(0+) = \delta_{iM}, \qquad k = 0, \cdots, M - i, \qquad \text{and} \qquad i = 0, \cdots, M. \tag{14}$$

We note that $i = M$ implies that $k = 0$. Also,

$$P_i(\tau) = \sum_{k=0}^{M-i} P_{ik}(\tau), \tag{15}$$

and if we sum (13) and (14) from $k = 0$ to $M - i$ we obtain (4) and (5).

It follows directly from (12), (13), and (15) that

$$\frac{dF}{d\tau} = \mu \sum_{i=1}^{M} iP_i(\tau) = \mu N(\tau). \tag{16}$$

This implies that the rate of change of the expected number of records removed from the MFS by the going-forward people is equal to the expected value at time τ of the rate at which change and deletion orders are received. If we let $C(\tau)$ denote the expected number of records removed from the MFS by the crash people in $(0, \tau]$, then

$$C(\tau) = M - N(\tau) - F(\tau). \tag{17}$$

From (7), (16), and (17), we obtain

$$\frac{dC}{d\tau} = \mu\nu[1 - P_0(\tau)] = \mu\nu \sum_{i=0}^{M} (1 - \delta_{i0})P_i(\tau), \tag{18}$$

which has an interpretation analogous to that of (16). We note that $dC/d\tau \approx \mu\nu$ in the region where $P_0(\tau)$ is small. Also, corresponding to the approximation $N(\tau) \approx L(\tau)$, where $L(\tau)$ is defined in (8), it follows from (16) that

$$F(\tau) \approx (M + \nu)(1 - e^{-\mu\tau}) - \mu\nu\tau, \tag{19}$$

since $F(0+) = 0$.

III. EXPECTED PASSAGE TIMES

We next turn to the calculation of the expected passage time from l to m records in the MFS. Now,[5] for a pure death process, with death rate μ_i, the passage time $\tau_{i,i-1}$ from state i to state $i - 1$ is exponentially distributed with density $\mu_i \exp(-\mu_i\tau)$. Hence, in particular, the expected value of $\tau_{i,i-1}$ is

$$E\tau_{i,i-1} = \frac{1}{\mu_i}. \tag{20}$$

The passage time from l to m records in the MFS is

$$\tau_{l,m} = \sum_{i=m+1}^{l} \tau_{i,i-1}, \tag{21}$$

and the random variables in the sum are independent. From (3), (20), and (21), we obtain

$$\mu E\tau_{l,m} = \sum_{i=m+1}^{l} \frac{1}{(\nu + i)} = \psi(l + \nu + 1) - \psi(m + \nu + 1), \qquad (22)$$

where ψ denotes the logarithmic derivative of the gamma function.[4]

We will make use of the asymptotic result

$$\psi(x) = \log x + O\left(\frac{1}{x}\right) \qquad \text{for} \quad x \gg 1. \qquad (23)$$

Hence, from (22), the expected time at which the MFS becomes empty is given by

$$\begin{aligned} \mu E\tau_{M,0} &= \psi(M + \nu + 1) - \psi(\nu + 1) \\ &\sim \log\left(\frac{M + \nu}{\nu}\right) \qquad \text{for} \quad \nu \gg 1. \end{aligned} \qquad (24)$$

Also, the expected time starting from l records until the MFS becomes empty is given by

$$\mu E\tau_{l,0} = \psi(l + \nu + 1) - \psi(\nu + 1) \sim \log\left(\frac{l + \nu}{\nu}\right) \qquad \text{for} \quad \nu \gg 1$$

$$\sim \frac{1}{\nu} \qquad \text{for} \quad \nu \gg 1, \qquad l \ll \nu. \qquad (25)$$

Next, the expected time at which the number of records in the MFS first reaches m is given by

$$\begin{aligned} \mu E\tau_{M,m} &= \psi(M + \nu + 1) - \psi(m + \nu + 1) \\ &\sim \log\left(\frac{M + \nu}{m + \nu}\right) \qquad \text{for} \quad \nu \gg 1. \end{aligned} \qquad (26)$$

But, from (8) to (10), the time τ_m at which the expected number of records in the MFS is equal to m satisfies

$$\mu\tau_m \sim \log\left(\frac{M + \nu}{m + \nu}\right)$$

$$\text{for} \quad M \gg 1, \qquad \nu \gg 1, \qquad m \gg \min(\sqrt{M}, \sqrt{\nu})/\sqrt{2\pi}. \qquad (27)$$

Hence, under the restrictions in (27), we have $E\tau_{M,m} \sim \tau_m$.

We now consider the final stage of the conversion by the crash force, starting from the time $\tau_{M,0}$ when the MFS becomes empty. Each of the n crash people is then busy with a record. We first assume that a person is removed from the crash force when he/she completes the conversion of the record on which he/she has been working. If j is the number of records remaining to be converted by the crash force, we then have a pure death process with $\mu_j = j\rho, j = 0, \cdots, n$. If τ_c denotes

the time at which the conversion by the crash force is completed, then, with the help of (2),

$$\mu E(\tau_c - \tau_{M,0}) = \mu \sum_{j=1}^{n} \frac{1}{\mu_j} = \frac{n}{\nu} \sum_{j=1}^{n} \frac{1}{j} = \frac{n}{\nu} [\psi(n+1) - \psi(1)]. \quad (28)$$

At the other extreme, we assume that the entire crash force works jointly on the remaining records, until the conversion of the last record is completed. This will give a lower bound on the expected time to complete the conversion of the final n records, since the crash force continues to work at its maximum rate. In this case the death rate is $\hat{\mu}_j = n\rho = \nu\mu$, $j = 1, \cdots, n$. If $\hat{\tau}_c$ denotes the corresponding time at which the conversion by the crash force is completed, then

$$\mu E(\hat{\tau}_c - \tau_{M,0}) = \mu \sum_{j=1}^{n} \frac{1}{\hat{\mu}_j} = \frac{n}{\nu}. \quad (29)$$

IV. CONCLUSION

In this paper we presented formulas that are employed in an economic model[1] used to estimate cost/benefits and labor force levels associated with the mechanization of transaction data bases. Because the formulas can be computed very rapidly on any modern computer, the model can be used as an interactive managerial decision tool to perform what-if studies in real time. Numerical results presented graphically are typical of two applications, namely, (*i*) the conversion of manual records in a service order processing operation of a Bell Operating Company business office, and (*ii*) the mechanization of equipment inventory records during the implementation of Trunks Integrated Record Keeping System.[1] We do not presume that the stochastic model is appropriate for all possible data base conversions, but evidence from field trials indicates that it does adequately model a certain class of Bell System mechanization problems.

V. ACKNOWLEDGMENTS

We would like to thank T. A. Bottomley for many helpful contributions concerning the definition of the problem, and J. B. Seery for writing the programs to obtain the numerical results presented in Figs. 1 and 2.

REFERENCES

1. W. W. Yale, unpublished work.
2. L. Kleinrock, *Queueing Systems, Volume I: Theory*, New York: John Wiley, 1975.
3. J. A. Morrison, unpublished work.
4. W. Magnus, F. Oberhettinger, and R. P. Soni, *Formulas and Theorems for the Special Functions of Mathematical Physics*, New York: Springer-Verlag, 1966.
5. J. Keilson, *Markov Chain Models—Rarity and Exponentiality*, Applied Mathematical Sciences, *28*, New York: Springer-Verlag, 1979, p. 21.